



# Adaptive Linear Models for Regression: improving prediction when population has changed

Charles Bouveyron, Julien Jacques

## ► To cite this version:

Charles Bouveyron, Julien Jacques. Adaptive Linear Models for Regression: improving prediction when population has changed. Pattern Recognition Letters, 2010, 31 (14), pp.2237-2247. hal-00305987v3

**HAL Id: hal-00305987**

**<https://hal.science/hal-00305987v3>**

Submitted on 30 Mar 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Adaptive linear models for regression: improving prediction when population has changed

Charles Bouveyron<sup>a</sup>, Julien Jacques<sup>b</sup>

<sup>a</sup>*Laboratoire SAMM, Université Paris I Panthéon-Sorbonne, Paris, France.*

<sup>b</sup>*Laboratoire Paul Painlevé, UMR CNRS 8524, Université Lille I, Lille, France.*

---

## Abstract

The general setting of regression analysis is to identify a relationship between a response variable  $Y$  and one or several explanatory variables  $X$  by using a learning sample. In a prediction framework, the main assumption for predicting  $Y$  on a new sample of observations is that the regression model  $Y = f(X) + \epsilon$  is still valid. Unfortunately, this assumption is not always true in practice and the model could have changed. We therefore propose to adapt the original regression model to the new sample by estimating a transformation between the original regression function  $f(X)$  and the new one  $f^*(X)$ . The main interest of the proposed adaptive models is to allow the build of a regression model for the new population with only a small number of observations using the knowledge on the reference population. The efficiency of this strategy is illustrated by applications on artificial and real datasets, including the modeling of the housing market in different U.S. cities. A package for the R software dedicated to the adaptive linear models is available on the author's web page.

*Key words:* regression adaptive, estimation, knowledge transfer, linear transformation models, housing market in different U.S. cities.

## 1. Introduction

The general setting of regression analysis is to identify a relationship (the regression model) between a response variable and one or several explanatory variables. Most of the works in regression analysis has focused on the nature of the regression model: linear model [1], generalized linear model [2] and non linear model [3]. We refer to [4] for a general survey on regression analysis.

### 1.1. *The problem of adapting a knowledge to a new situation*

In this paper, we are concerned with the following question: how to adapt an existing regression model to a new situation, for which the variables are identical (with a possible different probability density distribution) but where the relationship between response and explanatory variables could have changed? As a motivating example, our discussion will be centered on the following socio-economical application: a real-estate agency of the US East coast has to its disposal, through their long experience in this area, a regression model of the housing price versus several housing descriptive variables estimated using a large learning sample. To conquer new markets, this company plans to open several agencies in the West coast, and would use its regression model without having to spend a lot of time and money in collecting new housing market data for this area. Considering that the link between housing descriptive variables and housing price for the West and East coasts is, on the one hand, probably not the same but, on the other hand, not completely different, this work will consider a set of transformation models between both West and East coast regression models. This paper will therefore focus on transferring the knowledge on a reference population to a

25 new population by inferring the relationship between both regression models.  
26 Moreover, the exhibition of a link between both populations could be helpful  
27 for the interpretation of the modeled phenomenon.

## 28 1.2. Related works

29 To our knowledge, there have been only few contributions dealing with  
30 this original problem although it is very interesting and very frequent in  
31 practical applications. In the machine learning community, a related prob-  
32 lem is investigated under the keyword *Covariate Shift*. The covariate shift  
33 problem considers that the probability density of the new data is different  
34 from the learning data and the regression model is assumed to be conserved.  
35 Thus, if the regression model is exactly known, a change in the probability  
36 distribution of the explanatory variables is not a problem. Unfortunately,  
37 this is never the case in practice and the regression model estimated with the  
38 learning data could be very disappointing when applied to data with a dif-  
39 ferent probability distribution. Several recent works [5–9] have contributed  
40 to analyze this context. However, most of these works need to know the  
41 probability distribution of the new data or at least an estimation of this  
42 probability distribution. In practice, this is a difficult problem which re-  
43 quires a sufficiently large sample of observations. The focus of the present  
44 work is more general and does not assume that the relationship between ex-  
45 planatory and response variables is conserved from the learning data to the  
46 new data. In addition, the situation under review in this paper considers  
47 that only few learning data are available for the new situation, which is not  
48 enough to correctly estimate in practice their probability distribution. In su-  
49 pervised classification, a similar problem was studied in [10] on quantitative

50 variables and in [11] in the case of binary variables. The authors considered  
51 a model-based discriminant rule for classifying individuals from a population  
52 which differs from the learning one. For this, they introduced a family of  
53 linear models modeling the transformation between the reference population  
54 and the new population. An extension of this work to logistic regression was  
55 recently proposed in [12]. Finally, some other works cover the problematic  
56 of knowledge transfer in specific industrial contexts. For instance, [13] gives  
57 a good overview of solutions for model transfer in the field of Chemomet-  
58 rics. Among the proposed transfer models, the most used models are the  
59 piecewise direct standardization [14] and the neural network based nonlinear  
60 transformation [15]. Several works [16, 17] have also considered this problem  
61 in the field of semiconductor industry.

62 This paper is structured as follows. Section 2 formulates the problem of  
63 adapting an existing regression model to a new population and Section 3 in-  
64 troduces a family of transformation models to solve this problem. Inference  
65 and model selection procedures are discussed in Section 4. Section 5 provides  
66 a simulation study in a spline regression context and two real applications  
67 including the modeling of the housing market in different U.S. cities. Finally,  
68 some concluding remarks and future directions are discussed in Section 6.

## 69 **2. Problem formulation**

70 In this section, after having reminded the general framework of regression  
71 analysis, the problem of adapting an existing regression model to another  
72 population is formulated.

73 *2.1. Linear models for regression*

74 In regression analysis, the data  $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ , arising from  
 75 a population  $P$ , are assumed to be independent and identically distributed  
 76 samples of a couple of variables  $(\mathbf{X}, Y)$  with an unknown distribution. The  
 77 observations  $\mathbf{x}_j$ ,  $j = 1, \dots, n$ , are the values of the deterministic explanatory  
 78 variable  $\mathbf{X} = (X^{(1)}, \dots, X^{(p)})^t \in \mathbb{R}^p$  and the corresponding  $y_j$  are the real-  
 79 izations of the stochastic variable  $Y \in \mathbb{R}$ . A general data modeling problem  
 80 consists in identifying the relationship between the explanatory variable  $\mathbf{X}$   
 81 (known as well as covariate) and the response variable  $Y$  (or dependent vari-  
 82 able). Both standard parametric and non-parametric approaches consider  
 83 the following regression model:

$$Y = f(\mathbf{X}, \boldsymbol{\beta}) + \epsilon, \quad (1)$$

84 where the residuals  $\epsilon \sim \mathcal{N}(0, \sigma^2)$  are independent and where  $\boldsymbol{\beta}$  is the vec-  
 85 tor of regression parameters. This model is equivalent to the distributional  
 86 assumption that:

$$Y|\mathbf{X} \sim \mathcal{N}(f(\mathbf{X}, \boldsymbol{\beta}), \sigma^2),$$

87 where the regression function  $f(\mathbf{x}, \boldsymbol{\beta})$  is defined as the conditional expec-  
 88 tation  $E[Y|\mathbf{X} = \mathbf{x}]$ . Therefore, the only way to specify the link between  
 89 the response variable  $Y$  and the covariate  $\mathbf{X}$  is through the assumptions on  
 90  $f(\mathbf{x}, \boldsymbol{\beta})$ . In particular, parametric regression achieves this connection by as-  
 91 suming a specific form for  $f(\mathbf{x}, \boldsymbol{\beta})$ . The most common model is the linear  
 92 form (*cf.* [18]):

$$f(\mathbf{x}, \boldsymbol{\beta}) = \sum_{i=0}^d \beta_i \psi_i(\mathbf{x}), \quad (2)$$

93 where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)^t \in \mathbb{R}^{d+1}$  are the regression parameters,  $\psi_0(\mathbf{x}) = 1$   
 94 and  $(\psi_i)_{1 \leq i \leq d}$  is a basis of regression functions:

$$\psi_i : \mathbb{R}^p \rightarrow \mathbb{R},$$

95 which can be for instance the identity, polynomial functions, splines func-  
 96 tions [19] or wavelets [20]. We refer to [4] for a general survey. Let us notice  
 97 that the usual linear regression occurs when  $d = p$  and  $\psi_i(\mathbf{x}) = x^{(i)}$  for  
 98  $i = 1, \dots, d$ . The regression function (2) can be also written in a matrix  
 99 form as follows:

$$f(\mathbf{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^t \boldsymbol{\Psi}(\mathbf{x}), \quad (3)$$

100 where  $\boldsymbol{\Psi}(\mathbf{x}) = (1, \psi_1(\mathbf{x}), \dots, \psi_d(\mathbf{x}))^t$ .

## 101 2.2. How to adapt a regression model to another population?

102 Let us now assume that the regression function  $f$  has been estimated in  
 103 a preliminary study by using a sample  $S$  of population  $P$ , and that a new re-  
 104 gression model has to be adjusted on a new sample  $S^* = \{(\mathbf{x}_1^*, y_1^*), \dots, (\mathbf{x}_{n^*}^*, y_{n^*}^*)\}$ ,  
 105 measured on the same explanatory variables but arising from another pop-  
 106 ulation  $P^*$  ( $n^*$  is assumed to be small). The difference between  $P$  and  $P^*$   
 107 can be for instance geographical (as in the U.S. housing market application)  
 108 or temporal. However, the nature of both populations has to be similar to  
 109 match the purpose of this paper. The new regression model for  $P^*$  can be  
 110 classically written:

$$Y^* | \mathbf{X}^* \sim \mathcal{N}(f^*(\mathbf{X}^*, \boldsymbol{\beta}^*), \sigma^{*2}), \quad (4)$$

with

$$f^*(\mathbf{x}^*, \boldsymbol{\beta}^*) = \sum_{i=0}^{d^*} \beta_i^* \psi_i^*(\mathbf{x}^*) = \boldsymbol{\beta}^{*t} \boldsymbol{\Psi}^*(\mathbf{x}^*).$$

111 The aim of this work is therefore to define a link between the regression  
 112 functions  $f$  and  $f^*$ .

### 113 3. Adaptive linear models for regression

114 In this section, a link between the regression function of  $P$  and  $P^*$  is  
 115 exhibited, and a family of transformations is then introduced to solve the  
 116 problem of adapting an existing regression model of a reference population  $P$   
 117 to a new population  $P^*$ .

#### 118 3.1. The transformation model

119 In order to exhibit a link between both regression functions, we make the  
 120 following important assumptions.

121 *Assumption (A1).* First, we postulate that the number of basis functions  
 122 and the basis functions themselves are the same for both regression models  
 123 ( $d^* = d$  and  $\psi_i^* = \psi_i, \forall i = 1, \dots, d$ ), which is natural since the variables are  
 124 identical in both populations. The regression function of the population  $P^*$   
 125 can be therefore written:

$$f^*(\mathbf{x}^*, \boldsymbol{\beta}^*) = \boldsymbol{\beta}^{*t} \Psi(\mathbf{x}^*).$$

126 *Assumption (A2).* Second, we assume that the transformation between  $f$   
 127 and  $f^*$  applies only on the regression parameters. We therefore define  $\boldsymbol{\Lambda}$ ,  
 128 the transformation matrix between the regression parameters  $\boldsymbol{\beta}$  and  $\boldsymbol{\beta}^*$ , as  
 129 follows:

$$\boldsymbol{\beta}^* = \boldsymbol{\Lambda} \boldsymbol{\beta},$$



130 and this yields to the following expression of  $f^*$ :

$$f^*(\mathbf{x}^*, \mathbf{\Lambda}\boldsymbol{\beta}) = (\mathbf{\Lambda}\boldsymbol{\beta})^t \Psi(\mathbf{x}^*). \quad (5)$$

131 Given that the number of parameters to estimate in the transformation ma-  
 132 trix  $\mathbf{\Lambda}$  is  $(d+1) \times (d+1)$  and that the number of free parameters for learning  
 133 a new regression model directly from the sample  $S^*$  is  $(d+1)$ , the transfor-  
 134 mation model (5) is consequently highly over-parametrized. It is therefore  
 135 necessary to introduce some constraints on the transformation model such  
 136 that the number of free parameters to estimate is lower or equal to  $d$ .

137 *Assumption (A3).* Third, we assume that the relation between the response  
 138 variable and a specific covariate in the new population  $P^*$  only depends on  
 139 the relation between the response variable and the same covariate in the  
 140 population  $P$ . Thus, for  $i = 0, \dots, d$ , the regression parameter  $\beta_i^*$  only  
 141 depends on the regression parameter  $\beta_i$  and the matrix  $\mathbf{\Lambda}$  is consequently  
 142 diagonal. The transformation can be finally written in term of the regression  
 143 parameters of both models as follows:

$$\beta_i^* = \lambda_i \beta_i \quad \forall i = 0, \dots, d, \quad (6)$$

144 where  $\lambda_i \in \mathbb{R}$  is the  $i$ -th diagonal element of  $\mathbf{\Lambda}$ .

### 145 3.2. A family of transformation models

146 Since the aim of this study is to learn a regression model for  $P^*$  with only  
 147 few observations, we define in this section parsimonious models by imposing  
 148 some constraints on the transformation model (6). First, we allow some of the  
 149 parameters  $\lambda_i$  to be equal to 1 (in this case the regression parameters  $\beta_i^*$  are

150 equal to  $\beta_i$ ). Second, we allow as well some of the parameters  $\lambda_i$  to be equal  
 151 to a common value, *i.e.*  $\lambda_i = \lambda$  for given  $0 \leq i \leq d$ . The number of possible  
 152 models obtained with such a strategy is consequently very large (formally  
 153  $\sum_{m=0}^{d+1} \binom{d+1}{m} \times (1 + \sum_{l=2}^m \binom{l}{m})$ ) and it obviously impossible to consider all  
 154 these models in practice. These models, named adaptive linear models in the  
 155 sequel, are declined below into two families: specific transformation models  
 156 and prior-based transformation models.

### 157 3.2.1. Specific transformation models

158 We propose in this paragraph a family of 7 transformation models, se-  
 159 lected on parsimony and interpretability criteria, ranging from the most com-  
 160 plex model (hereafter  $M0$ ) to the simplest one (hereafter  $M6$ ):

- 161 • Model  $M0$ :  $\beta_0^* = \lambda_0 \beta_0$  and  $\beta_i^* = \lambda_i \beta_i$ , for  $i = 1, \dots, d$ . This model is  
 162 the most complex model of transformation between the populations  $P$   
 163 and  $P^*$ . It is equivalent to learning a new regression model from the  
 164 sample  $S^*$ , since there is no constraint on the  $d + 1$  parameters  $\beta_i^*$   
 165 ( $i = 0, \dots, d$ ), and the number of free parameters in  $\mathbf{\Lambda}$  is consequently  
 166  $d + 1$  as well.
- 167 • Model  $M1$ :  $\beta_0^* = \beta_0$  and  $\beta_i^* = \lambda_i \beta_i$  for  $i = 1, \dots, d$ . This model assumes  
 168 that both regression models have the same intercept  $\beta_0$ .
- 169 • Model  $M2$ :  $\beta_0^* = \lambda_0 \beta_0$  and  $\beta_i^* = \lambda \beta_i$  for  $i = 1, \dots, d$ . This model assumes  
 170 that the intercept of both regression models differ by the scalar  $\lambda_0$  and  
 171 all the other regression parameters differ by the same scalar  $\lambda$ .
- 172 • Model  $M3$ :  $\beta_0^* = \lambda \beta_0$  and  $\beta_i^* = \lambda \beta_i$  for  $i = 1, \dots, d$ . This model assumes

Table 1: Complexity (number of parameters  $\nu$ ) of the transformation models.

| Model                        | $M0$               | $M1$               | $M2$               | $M3$             | $M4$             | $M5$               | $M6$      |
|------------------------------|--------------------|--------------------|--------------------|------------------|------------------|--------------------|-----------|
| $\beta_0^*$ is assumed to be | $\lambda_0\beta_0$ | $\beta_0$          | $\lambda_0\beta_0$ | $\lambda\beta_0$ | $\beta_0$        | $\lambda_0\beta_0$ | $\beta_0$ |
| $\beta_i^*$ is assumed to be | $\lambda_i\beta_i$ | $\lambda_i\beta_i$ | $\lambda\beta_i$   | $\lambda\beta_i$ | $\lambda\beta_i$ | $\beta_i$          | $\beta_i$ |
| Nb. of parameters $\nu$      | $d+1$              | $d$                | 2                  | 1                | 1                | 1                  | 0         |

that all the regression parameters of both regression models differ by the same scalar  $\lambda$ .

- Model  $M4$ :  $\beta_0^* = \beta_0$  and  $\beta_i^* = \lambda\beta_i$  for  $i = 1, \dots, d$ . This model assumes that both regression models have the same intercept  $\beta_0$  and all the other regression parameters differ by the same scalar  $\lambda$ .

- Model  $M5$ :  $\beta_0^* = \lambda_0\beta_0$  and  $\beta_i^* = \beta_i$  for  $i = 1, \dots, d$ . This model assumes that both regression models have the same parameters except the intercept.

- Model  $M6$ :  $\beta_0^* = \beta_0$  and  $\beta_i^* = \beta_i$  for  $i = 1, \dots, d$ . This model assumes that both populations  $P$  and  $P^*$  have the same regression model.

The numbers of parameters to estimate for these transformation models are presented in Table 1. The choice of this family is arbitrary and motivated by the will of the authors to treat similarly all the covariates in this general discussion. However, in practical applications, we encourage the practitioner to consider some additional transformation models specifically designed to his application and motivated by his prior knowledge on the subject. This is discussed in the next section.

### 190 3.2.2. Prior-based transformation models

191 Although only seven pragmatic transformation models have been pre-  
 192 sented in the previous section, some other transformation models could be  
 193 considered as well, for which the complexity (in number of parameters) will  
 194 be intermediate between the  $M1$  complexity ( $d$ ) and the  $M2$  complexity  
 195 (2). Indeed, the practitioner could have in some specific cases to use in-  
 196 termediate transformation models suggested by some prior informations on  
 197 the covariates, which leads to impose specific constraints on parameters  $\lambda_i$   
 198 for given  $i \in \{1, \dots, d\}$ . For instance, let us consider the specific transfor-  
 199 mation matrix  $\mathbf{\Lambda} = \text{diag}(\lambda_0, \lambda_1, \lambda, \dots, \lambda)$  where  $\text{diag}(\lambda_0, \lambda_1, \lambda, \dots, \lambda)$  is the  
 200  $(d+1) \times (d+1)$  diagonal matrix having  $\{\lambda_0, \lambda_1, \lambda, \dots, \lambda\}$  on its diagonal.  
 201 This model assumes that the regression parameters  $\beta_i, i = 2, \dots, d$  are trans-  
 202 formed in the same manner whereas the intercept and  $\beta_1$  are not.

## 203 4. Estimation procedure and model selection

204 The estimation procedure associated with the adaptive linear models,  
 205 proposed in the previous section, is made of two main steps corresponding to  
 206 the estimation of the regression parameters of the population  $P$  and to the  
 207 estimation of the transformation parameters using samples of the popula-  
 208 tion  $P^*$ . The regression parameters of  $P^*$  are then obtained by plug-in. The  
 209 ordinary least square (OLS) method is used, but we present in this paper  
 210 the equivalent maximum likelihood estimation method in order to compute  
 211 penalized likelihood model selection criteria for model selection.

212 *4.1. Estimation of the regression parameters*

213 Let us consider a data set of inputs  $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  with corresponding  
 214 response values grouped in a column vector  $\mathbf{y} = (y_1, \dots, y_n)^t$ . Under the  
 215 assumptions of the model (2), the log-likelihood of  $\mathbf{y}$  given  $\mathbf{x}$ ,  $\boldsymbol{\beta}$  and  $\sigma^2$  is:

$$\ln l(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}, \sigma^2) = -n \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{j=1}^n (y_j - \boldsymbol{\beta}^t \Psi(\mathbf{x}_j))^2. \quad (7)$$

216 Maximizing the log-likelihood according to  $\boldsymbol{\beta}$  is equivalent to minimizing  
 217  $\sum_{j=1}^n (y_j - \boldsymbol{\beta}^t \Psi(\mathbf{x}_j))^2$  and thus the maximum likelihood estimator is equiv-  
 218 alent to the ordinary least square estimator:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{OLS} &= (\boldsymbol{\Psi}^t \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^t \mathbf{y}, \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{j=1}^n (y_j - \boldsymbol{\beta}^t \Psi(\mathbf{x}_j))^2. \end{aligned}$$

219 where  $\boldsymbol{\Psi}$  is a  $(n) \times (d+1)$  matrix formed by the row vector  $\Psi(\mathbf{x}_j)^t$  ( $1 \leq j \leq n$ ).

220 *4.2. Estimation of the transformation parameters*

221 For this second step, it is assumed that  $\boldsymbol{\beta}$  is known (in fact it is estimated  
 222 in the previous step). As previously noticed, the full model  $M0$  corresponds  
 223 to a completely new regression model adjusted on the sample  $S^*$ . Similarly,  
 224 the model  $M6$ , which assumes no transformation between  $P$  and  $P^*$ , does  
 225 not require the estimation of any transformation parameters. Let us consider  
 226 now a sample  $\mathbf{x}^* = \{\mathbf{x}_1^*, \dots, \mathbf{x}_{n^*}^*\}$  drawn from  $P^*$  with corresponding response  
 227 values  $\mathbf{y}^* = (y_1^*, \dots, y_{n^*}^*)^t$ . By replacing  $\boldsymbol{\beta}^* = \mathbf{\Lambda} \boldsymbol{\beta}$  in (7), the log-likelihood  
 228 of model (4) is:

$$\ln l(\mathbf{y}^*; \mathbf{x}^*, \mathbf{\Lambda}, \sigma^2) = -n^* \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{j=1}^{n^*} (y_j^* - \boldsymbol{\beta}^t \mathbf{\Lambda}^t \Psi(\mathbf{x}_j^*))^2. \quad (8)$$

229 This log-likelihood must be maximized according to the transformation ma-  
 230 trix  $\mathbf{\Lambda}$ , what leads to the OLS estimator:

$$\hat{\mathbf{\Lambda}}^{OLS} = \underset{\mathbf{\Lambda} \in \mathcal{D}}{\operatorname{argmin}} \sum_{j=1}^{n^*} (y_j^* - (\mathbf{\Lambda}\boldsymbol{\beta})^t \Psi(\mathbf{x}_j^*))^2, \quad (9)$$

231 where  $\mathcal{D}$  is a set of diagonal matrices depending on the model of transforma-  
 232 tion at hand. For instance, with the model  $M3$ , this set is  $\mathcal{D} = \{\lambda I_{d+1}, \lambda \in \mathbb{R}\}$   
 233 where  $I_{d+1}$  is the identity matrix of size  $d + 1$ .

#### 234 4.2.1. Specific transformation models

235 Least square estimators of the specific models  $M1$  to  $M5$  are derived  
 236 below.

237 *Model  $M_1$ .* As the transformation matrix is  $\mathbf{\Lambda} = \operatorname{diag}(1, \lambda_1, \dots, \lambda_d)$ , the log-  
 238 likelihood (8) can be written

$$\ln l(\mathbf{y}^*; \mathbf{x}^*, \mathbf{\Lambda}, \sigma^2) = -n^* \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{j=1}^{n^*} (y_j^* - \beta_0 - \boldsymbol{\beta}_{\sim 1}^t \mathbf{\Lambda}_{\sim 1}^t \Psi_{\sim 1}(\mathbf{x}_j^*))^2$$

239 where  $\mathbf{\Lambda}_{\sim k}$  and  $\boldsymbol{\beta}_{\sim k}$  correspond respectively to  $\mathbf{\Lambda}$  and  $\boldsymbol{\beta}$  without the  $k$ -th  
 240 row. This maximization is therefore similar to the maximization of (7) and  
 241 leads to the following estimator of  $\mathbf{\Lambda}_{\sim 1} = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ :

$$\hat{\mathbf{\Lambda}}_{\sim 1}^{OLS} = (\boldsymbol{\Psi}_{\sim 1}^{*t} \boldsymbol{\beta}_{\sim 1} \boldsymbol{\beta}_{\sim 1}^t \boldsymbol{\Psi}_{\sim 1}^*)^{-1} \boldsymbol{\beta}_{\sim 1}^t \boldsymbol{\Psi}_{\sim 1}^* (\mathbf{y}^* - \beta_0)$$

242 where  $\boldsymbol{\Psi}^*$  is a  $(n^*) \times (d + 1)$  matrix formed by the row vector  $\Psi(\mathbf{x}_j^*)^t$  ( $1 \leq$   
 243  $j \leq n^*$ ).

244 *Model  $M_2$ .* The transformation matrix has in this case the form  $\mathbf{\Lambda} = \operatorname{diag}(\lambda_0, \lambda, \dots, \lambda)$ .  
 245 The maximization according to  $\mathbf{\Lambda}$  of the following log-likelihood:

$$\ln l(\mathbf{y}^*; \mathbf{x}^*, \mathbf{\Lambda}, \sigma^2) = -n^* \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{j=1}^{n^*} (y_j^* - \beta_0 \lambda_0 - \boldsymbol{\beta}_{\sim 1}^t \mathbf{\Lambda}_{\sim 1}^t \Psi_{\sim 1}(\mathbf{x}_j^*))^2$$

246 leads to the estimator of  $\mathbf{\Lambda}_{M_2} = (\lambda_0, \lambda)^t$ :

$$\hat{\mathbf{\Lambda}}_{M_2}^{OLS} = (\mathbf{Q}^t \mathbf{Q})^{-1} \mathbf{Q}^t \mathbf{y}^*,$$

247 where

$$\mathbf{Q} = \begin{pmatrix} \beta_0 & \sum_{i=1}^d \beta_i \psi_i(\mathbf{x}_1^*) \\ & \vdots \\ \beta_0 & \sum_{i=1}^d \beta_i \psi_i(\mathbf{x}_n^*) \end{pmatrix}.$$

248 *Model  $M_3$ .* For this model, the transformation matrix is formed by only one  
 249 real parameter and  $\mathbf{\Lambda} = \text{diag}(\lambda, \lambda, \dots, \lambda)$ . The maximization of the log-  
 250 likelihood according to  $\lambda$  leads to the following estimator:

$$\hat{\lambda}^{OLS} = (\mathbf{\Psi}^{*t} \boldsymbol{\beta} \boldsymbol{\beta}^t \mathbf{\Psi}^*)^{-1} \boldsymbol{\beta}^t \mathbf{\Psi}^* \mathbf{y}^*.$$

251 *Model  $M_4$ .* In this case, the transformation matrix is formed by a constant  
 252 and a unique transformation parameter  $\lambda$ . The transformation matrix has  
 253 therefore the form  $\mathbf{\Lambda} = \text{diag}(1, \lambda, \dots, \lambda)$  and the corresponding estimator of  
 254  $\lambda$  is:

$$\hat{\lambda}^{OLS} = (\mathbf{\Psi}_{\sim 1}^{*t} \boldsymbol{\beta}_{\sim 1} \boldsymbol{\beta}_{\sim 1}^t \mathbf{\Psi}_{\sim 1}^*)^{-1} \boldsymbol{\beta}_{\sim 1}^t \mathbf{\Psi}_{\sim 1}^* (\mathbf{y}^* - \beta_0).$$

255 *Model  $M_5$ .* For this model, the transformation matrix is  $\mathbf{\Lambda} = \text{diag}(\lambda_0, 1, \dots, 1)$   
 256 and the estimator of  $\lambda_0$  is:

$$\hat{\lambda}_0^{OLS} = \frac{1}{n^* \beta_0} \sum_{j=1}^{n^*} [y_j^* - \sum_{i=1}^d \beta_i \psi_i(\mathbf{x}_j^*)].$$

#### 257 4.2.2. Prior-based transformation models

As previously discussed, the practitioner may prefer in some cases to use some particular transformation models suggested by some prior informations.

A generic transformation model including all possible particular transformation models and the corresponding estimators is described below. In the sequel, the subscripts  $\gamma_j$  will be associated with regression parameters of the new population to estimate using the relation  $\beta_{\gamma_j}^* = \lambda_{\gamma_j} \beta_{\gamma_j}$  with  $j = 1, \dots, q$  and  $\gamma_j \in \{0, \dots, d\}$ . In the same manner, the subscripts  $\bar{\gamma}_j$  will be associated with regression parameters of the new population which are similar to the original population parameters, *i.e.*  $\beta_{\bar{\gamma}_j}^* = \beta_{\bar{\gamma}_j}$  with  $j = 1, \dots, p - q$  and  $\bar{\gamma}_j \in 0, \dots, d$ . The regression model for the new population can be written as follows:

$$Y = \mathbf{Q}\mathbf{\Lambda}_q + \bar{\mathbf{Q}}\mathbf{1}_{p-q} + \epsilon,$$

258 where:

259     •  $\mathbf{\Lambda}_q = (\lambda_{\gamma_1}, \dots, \lambda_{\gamma_q})^t,$

260     •  $\mathbf{Q} = \begin{pmatrix} \beta_{\gamma_1} \psi_{\gamma_1}(x_1) & \cdots & \beta_{\gamma_q} \psi_{\gamma_q}(x_1) \\ \vdots & & \vdots \\ \beta_{\gamma_1} \psi_{\gamma_1}(x_n) & \cdots & \beta_{\gamma_q} \psi_{\gamma_q}(x_n) \end{pmatrix},$

261     •  $\bar{\mathbf{Q}} = \begin{pmatrix} \beta_{\bar{\gamma}_1} \psi_{\bar{\gamma}_1}(x_1) & \cdots & \beta_{\bar{\gamma}_q} \psi_{\bar{\gamma}_q}(x_1) \\ \vdots & & \vdots \\ \beta_{\bar{\gamma}_1} \psi_{\bar{\gamma}_1}(x_n) & \cdots & \beta_{\bar{\gamma}_q} \psi_{\bar{\gamma}_q}(x_n) \end{pmatrix},$

262     •  $\mathbf{1}_{p-q}$  is the unity vector of dimension  $p - q$ .

Consequently the maximum likelihood estimator of  $\mathbf{\Lambda}_q$  is

$$\hat{\mathbf{\Lambda}}_q^{OLS} = (\mathbf{Q}^t \mathbf{Q})^{-1} \mathbf{Q}^t (\mathbf{y} - \bar{\mathbf{Q}} \mathbf{1}_{p-q}).$$



### 263 4.3. Full and profile likelihood estimation

264 In this work, a reference regression model on the population  $P$  is as-  
 265 sumed to be known and is transformed in a new regression model adapted to  
 266 a new population  $P^*$  by estimating a transformation between both reference  
 267 and new populations. However, the regression parameters of the reference  
 268 model are in practice never known but only estimated from a given sam-  
 269 ple  $S$ . Therefore, starting from this estimation for inferring the new regres-  
 270 sion model could be disappointing in some cases, particularly when the size  
 271  $n$  of  $S$  is not large too. As both populations  $P$  and  $P^*$  are assumed to be  
 272 linked, it could be interesting to use both samples  $S$  and  $S^*$  for improving the  
 273 estimation of the regression parameter  $\beta$  as well. But, as the parameters  $\beta$   
 274 and  $\Lambda$  appear as a product in the regression equation (5) for the sample  $S^*$ ,  
 275 the full likelihood estimation of  $(\beta, \Lambda)$  can not be achieved directly and is  
 276 replaced by a profile likelihood estimation procedure. Starting from a ini-  
 277 tialization value  $\beta^{(0)}$  of  $\beta$ , the following two steps iteratively alternate until  
 278 the growth of the model likelihood is lower than a given threshold. At the  
 279 iteration  $q$ :

- 280 1. Compute the estimation  $\hat{\Lambda}^{(q)}$  of  $\Lambda$  given a current value of  $\hat{\beta}^{(q-1)}$  (this  
 281 step was the purpose of the previous section),
- 282 2. Compute the estimation  $\hat{\beta}^{(q)}$  of  $\beta$  given the estimation of  $\hat{\Lambda}^{(q)}$  obtained  
 283 in the previous step.

284 For a given estimation  $\hat{\Lambda}^{(q)}$  of  $\Lambda = \text{diag}(\lambda_0, \lambda_1, \dots, \lambda_d)$ , the estimation  
 285 of  $\beta$  consists in maximizing the log-likelihood of the considered regression  
 286 model (2) for the sample  $S$  and the log-likelihood of the same model in  
 287 which the regression function  $\psi_i$  are multiplied by  $\hat{\lambda}_i^{(q)}$  for the sample  $S^*$ . By

introducing  $\tilde{\mathbf{y}} = (y_1, \dots, y_n, y_1^*, \dots, y_{n^*}^*)^t$  and  $\tilde{\mathbf{\Psi}}$  the  $(n + n^*) \times (d + 1)$  matrix defined as follows:

$$\tilde{\mathbf{\Psi}} = \begin{pmatrix} \psi_0(\mathbf{x}_1) & \cdots & \psi_d(\mathbf{x}_1) \\ \vdots & & \vdots \\ \psi_0(\mathbf{x}_n) & \cdots & \psi_d(\mathbf{x}_n) \\ \hat{\lambda}_0^{(q)} \psi_0(\mathbf{x}_1^*) & \cdots & \hat{\lambda}_d^{(q)} \psi_d(\mathbf{x}_1^*) \\ \vdots & & \vdots \\ \hat{\lambda}_0^{(q)} \psi_0(\mathbf{x}_n^*) & \cdots & \hat{\lambda}_d^{(q)} \psi_d(\mathbf{x}_n^*) \end{pmatrix},$$

the estimator of  $\boldsymbol{\beta}$  given  $\hat{\mathbf{\Lambda}}^{(q)}$  is:

$$\hat{\boldsymbol{\beta}}^{OLS} = (\tilde{\mathbf{\Psi}}^t \tilde{\mathbf{\Psi}})^{-1} \tilde{\mathbf{\Psi}}^t \tilde{\mathbf{y}}.$$

#### 4.4. Assumption validation and model selection

In regression analysis, there is two indispensable steps: validation of the model assumptions and selection of the regression model.

*Assumption validation.* An important step in regression analysis is the validation of the linear model assumptions: independence and homoscedasticity of the residuals, linearity of the regression. In this context, several statistical tests have been defined, see for instance [4], and the practitioner would have to validate the linear model assumptions for the selected regression model as usually. In this paper the regression model for the population  $P$  is known and the estimation of the regression model for another population  $P^*$  is investigated, and it would be natural to test the equality of both regression models [21]. Unfortunately, this can not be achieved easily since there are too few available data in  $S^*$  to efficiently estimate the regression model on

304  $P^*$ . Nevertheless, the case of equality of the populations  $P$  and  $P^*$  is con-  
 305 sidered by the model  $M6$ , and a model selection procedure, described in the  
 306 next section, is carried out in place of the regression equality test.

307 *Model selection.* The second important step is the selection of the most ap-  
 308 propriate model of transformation between the populations  $P$  and  $P^*$ . We  
 309 propose to use three well-known criteria. The reader interested in a com-  
 310 parison of the respective performance of these three criteria could refer for  
 311 instance to [19]. The first criterion is the PRESS criterion [22], representing  
 312 the sum of squared prediction errors computed on a cross-validation scheme,  
 313 which is defined by:

$$PRESS = \frac{1}{n^*} \sum_{j=1}^{n^*} \|y_j^* - \hat{y}_j^{*-j}\|^2$$

314 where  $\hat{y}_j^{*-j}$  is the prediction of  $y_j^*$  obtained by the regression model estimated  
 315 without using the  $j$ -th individual  $y_j^*$  of the sample  $S^*$ . This criterion is  
 316 one of the most often used for model selection in regression analysis, and  
 317 we encourage its use when its computation is numerically feasible. Both  
 318 following penalized likelihood criteria are less computationally heavy. They  
 319 consist of selecting the models leading to the highest likelihood but penalizing  
 320 those which have a large number of parameters. The Bayesian Information  
 321 Criterion (BIC, [23]) is defined by:

$$BIC = -2 \ln \ell + \nu \ln n^*,$$

322 where  $\ell$  is the maximum likelihood value and  $\nu$  is the number of estimated  
 323 parameters (see Table 1). With the same notations, the Akaike Information

324 Criterion (AIC, [24]) penalized the log-likelihood by  $2\nu$ . For all these three  
 325 criteria, the most adapted model is the one with the smallest criterion value.

## 326 5. Experimental results

327 In this section, experimental results on artificial and real data illustrate  
 328 the main features of the adaptive linear models.

### 329 5.1. Simulation study

330 This first experiment aims to evaluate the ability of the adaptive linear  
 331 models, introduced in Section 3, to find the transformation between popula-  
 332 tions  $P$  and  $P^*$  as well as the ability of the model selection criteria to select  
 333 the most appropriate transformation model.

334 *Experimental setup.* Firstly, a one-dimensional regression model was gener-  
 335 ated for the reference population  $P$  on a basis of natural cubic Splines with  
 336 5 degrees of freedom. Then, a regression model was built for the new popu-  
 337 lation  $P^*$  from the model of  $P$  by multiplying the regression parameters of  $P$   
 338 by a given transformation matrix  $\mathbf{\Lambda}$ . Since it is difficult to report here numer-  
 339 ical experiments for all existing transformation models, results are presented  
 340 for only one transformation model: the model M2. Similar results could be  
 341 obtained for the other transformation models. The true regression model  
 342 for  $P$  is  $y = \sin(x) + \sin(2x) + \log(1 + x)$ , for  $x \in [0, \pi]$ , and the specific  
 343 transformation matrix  $\mathbf{\Lambda} = \text{diag}(1.5, 2, 2, 2, 2, 2)$  was chosen for generating  
 344 the regression model of  $P^*$ . The size  $n$  of the sample  $S$  was fixed to 1000.  
 345 In order to compare the performance of the different transformation mod-  
 346 els, some observations for population  $P^*$  were simulated from its regression

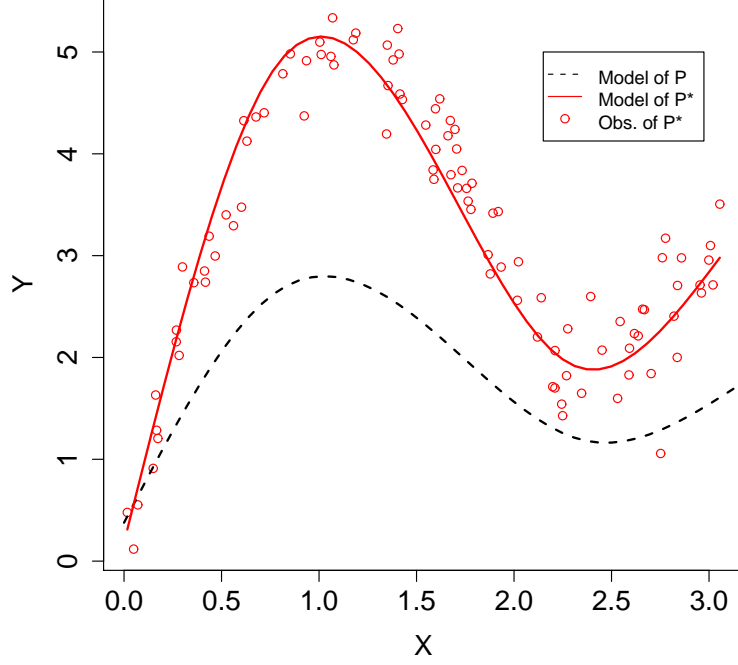


Figure 1: Regression models of the populations  $P$  and  $P^*$  and simulated observations of population  $P^*$ : the model of  $P$  was estimated on a basis of cubic Spline functions with 5 degrees of freedom and the model of  $P^*$  was obtained from the model of  $P$  by multiplying its parameters by  $(1.5, 2, 2, 2, 2, 2)$ .

347 model. These observations were simulated with an additive Gaussian noise  
348  $\epsilon \sim \mathcal{N}(0, 0.3)$ . Figure 1 shows the regression models for both populations  
349  $P$  and  $P^*$  as well as 100 observations simulated from the regression model  
350 of  $P^*$ . The simulated observations of population  $P^*$  were used in the ex-  
351 periment by the different linear transformation models for estimating the  
352 transformation between  $P$  and  $P^*$ . The values of the three model selection  
353 criteria, presented in Section 4.4, were computed for each model to verify

354 their ability to find the most appropriate transformation model. Finally, the  
 355 protocol described above was applied for different dataset sizes ranging from  
 356 25 to 1000 observations for studying the effect of the learning dataset size on  
 357 the prediction ability of the different models. The experiments were repeated  
 358 50 times in order to average the results.

359 *Experimental results.* Table 2 presents the numerical evaluation of the ability  
 360 of the adaptive linear models M0, M1, M2, M3, M4 and M5 to estimate the  
 361 transformation parameters and of the ability of the model selection criteria  
 362 to find the most appropriate transformation model. The first and the sec-  
 363 ond columns of Table 2 respectively indicate the size of the learning dataset  
 364 and the name of the used transformation model. The third, fourth and fifth  
 365 columns respectively give the values of the model selection criteria PRESS,  
 366 BIC and AIC associated to each model. Finally, the sixth column provides  
 367 the mean square error (MSE) computed on a test dataset different from the  
 368 learning set. The bold numbers of the table correspond to the “best values”  
 369 of each column for a given dataset size (let us remind that for the three model  
 370 selection criteria, the most appropriate model is the one associated with the  
 371 smallest value). On the one hand, it appears clearly that both PRESS, BIC  
 372 and AIC select the transformation model M2 as the most appropriate for  
 373 modeling the transformation between  $P$  and  $P^*$  and that corresponds to the  
 374 truth. The first conclusion is that these three criteria are well suited to select  
 375 the transformation model in such a case. On the other hand, it can be no-  
 376 ticed that the model M0, which corresponds to the usual OLS model on  $P^*$ ,  
 377 is very sensitive to the size of the dataset used for learning whereas the adap-  
 378 tive linear models M1 to M5 are less sensitive. Furthermore, the model M0

Table 2: Evaluation of the model selection and of the parameter estimation on data simulated according to the model M2 on a basis of cubic Spline functions for different dataset sizes: PRESS, BIC, AIC and MSE values are per point, and the MSE value was computed on a test dataset.

| $n^*$ | Model | PRESS        | BIC          | AIC          | MSE          |
|-------|-------|--------------|--------------|--------------|--------------|
| 25    | M0    | 24283.92     | 16.326       | 16.033       | 199.827      |
|       | M1    | 0.131        | 0.902        | 0.658        | 0.109        |
|       | M2    | <b>0.109</b> | <b>0.669</b> | <b>0.571</b> | <b>0.094</b> |
|       | M3    | 0.128        | 0.796        | 0.748        | 0.119        |
|       | M4    | 0.192        | 1.241        | 1.192        | 0.162        |
|       | M5    | 0.597        | 2.340        | 2.291        | 0.584        |
| 50    | M0    | 19196.07     | 16.209       | 15.979       | 51.884       |
|       | M1    | 0.098        | 0.669        | 0.478        | 0.103        |
|       | M2    | <b>0.091</b> | <b>0.498</b> | <b>0.421</b> | <b>0.096</b> |
|       | M3    | 0.111        | 0.661        | 0.623        | 0.119        |
|       | M4    | 0.157        | 1.042        | 1.004        | 0.163        |
|       | M5    | 0.525        | 2.220        | 2.182        | 0.545        |
| 100   | M0    | 1754.953     | 8.800        | 8.644        | 41.239       |
|       | M1    | 0.096        | 0.614        | 0.484        | 0.091        |
|       | M2    | <b>0.093</b> | <b>0.509</b> | <b>0.456</b> | <b>0.089</b> |
|       | M3    | 0.115        | 0.699        | 0.673        | 0.109        |
|       | M4    | 0.172        | 1.128        | 1.102        | 0.157        |
|       | M5    | 0.455        | 2.072        | 2.046        | 0.511        |
| 250   | M0    | 522.120      | 5.512        | 5.427        | 24.329       |
|       | M1    | 0.090        | 0.504        | 0.434        | 0.090        |
|       | M2    | <b>0.089</b> | <b>0.450</b> | <b>0.422</b> | <b>0.089</b> |
|       | M3    | 0.116        | 0.704        | 0.690        | 0.111        |
|       | M4    | 0.172        | 1.135        | 1.121        | 0.161        |
|       | M5    | 0.467        | 2.089        | 2.075        | 0.534        |
| 500   | M0    | 270.574      | 5.034        | 5.004        | 6.633        |
|       | M1    | 0.092        | 0.495        | 0.453        | 0.091        |
|       | M2    | <b>0.091</b> | <b>0.463</b> | <b>0.446</b> | <b>0.090</b> |
|       | M3    | 0.116        | 0.698        | 0.689        | 0.113        |
|       | M4    | 0.167        | 1.090        | 1.082        | 0.155        |
|       | M5    | 0.463        | 2.075        | 2.067        | 0.501        |
| 1000  | M0    | 184.00       | 4.669        | 4.618        | 3.519        |
|       | M1    | 0.089        | 0.450        | 0.425        | 0.091        |
|       | M2    | <b>0.089</b> | <b>0.432</b> | <b>0.422</b> | <b>0.090</b> |
|       | M3    | 0.113        | 0.669        | 0.665        | 0.112        |
|       | M4    | 0.168        | 1.093        | 1.088        | 0.156        |
|       | M5    | 0.453        | 2.051        | 2.046        | 0.501        |

379 gives disappointing estimations for all dataset sizes whereas the other mod-  
 380 els, which are more parsimonious and which benefit from the knowledge on  
 381  $P$ , give satisfying results for a large range of dataset sizes. Figure 2 shows the  
 382 estimated regression model of the population  $P^*$  for the six studied models.  
 383 These estimations were obtained with a learning dataset of 100 observations.  
 384 As it could be expected, the M0 estimation is very far away from the actual  
 385 model and the models M1, M2 and M3 give very good estimations of the  
 386 regression model. The effect of the constraints on the models can also be  
 387 observed on this figure. For instance, the model M5 is not flexible enough  
 388 to correctly estimate the transformation and this is due to the fact that it  
 389 assumes that only the intercept is modified. To summarize, this experiment  
 390 has shown that the adaptive linear models, proposed in the present paper,  
 391 are able to correctly estimate a transformation between two populations with  
 392 non-linear regression models and that even in situations where the number of  
 393 observations of  $P^*$  is limited. This study has also highlighted that either the  
 394 cross-validated PRESS criterion and information criteria BIC and AIC are  
 395 adapted to select the most appropriate model among the 7 adaptive linear  
 396 models.

## 397 5.2. Real data study: Growth of *Tetrahymena* cells

398 A biological dataset is considered here to highlight the ability of our  
 399 approach to deal with real data.

400 *The data.* The *hellung* dataset <sup>1</sup>, collected by P. Hellung-Larsen, reports the  
 401 growth conditions of *Tetrahymena* cells. The data arise from two groups of

---

<sup>1</sup>The hellung dataset is available in the ISwR package for R.



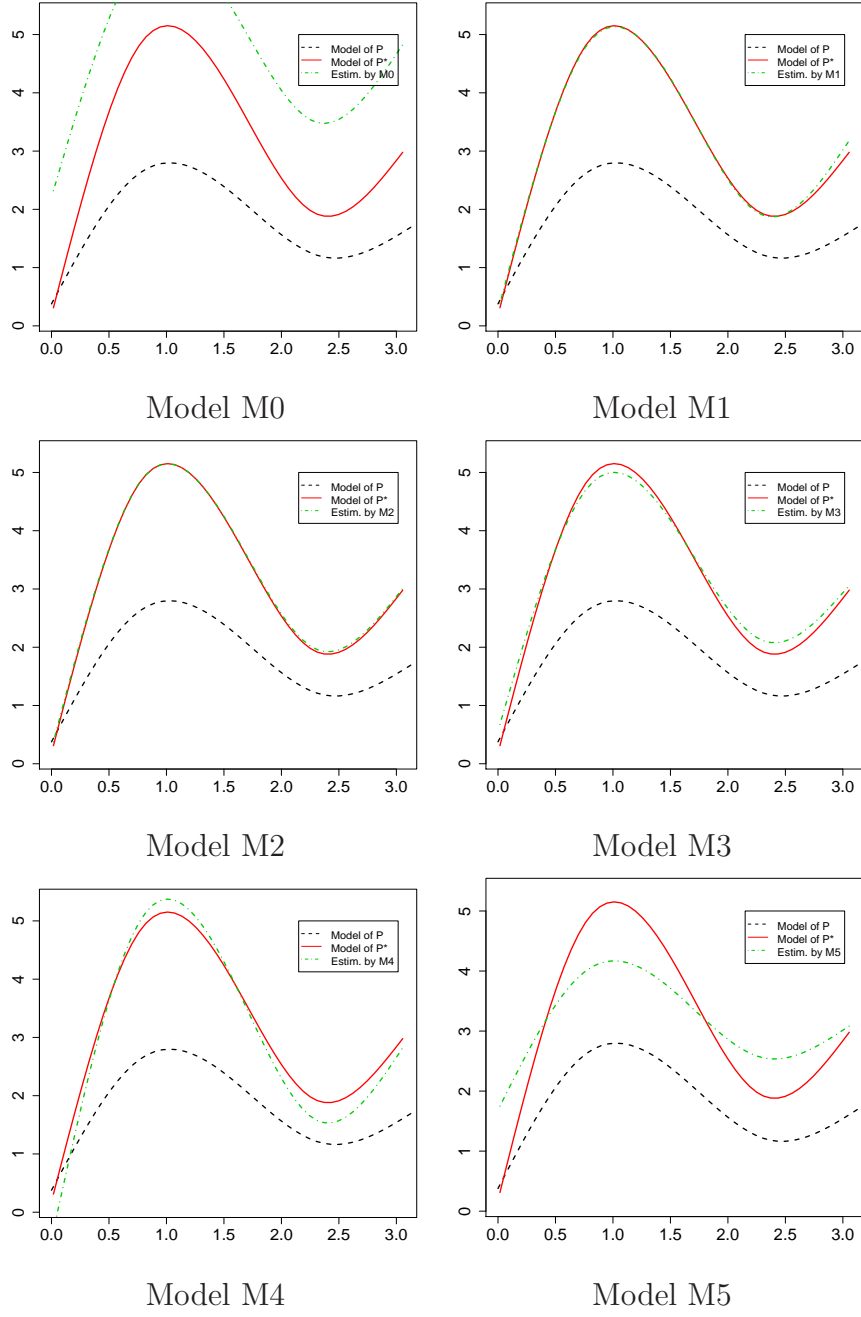


Figure 2: Parameter estimation with the different linear transformation models on data simulated according to the transformation model M2 on a basis of cubic Spline functions. These estimations were computed with a dataset of 100 observations.

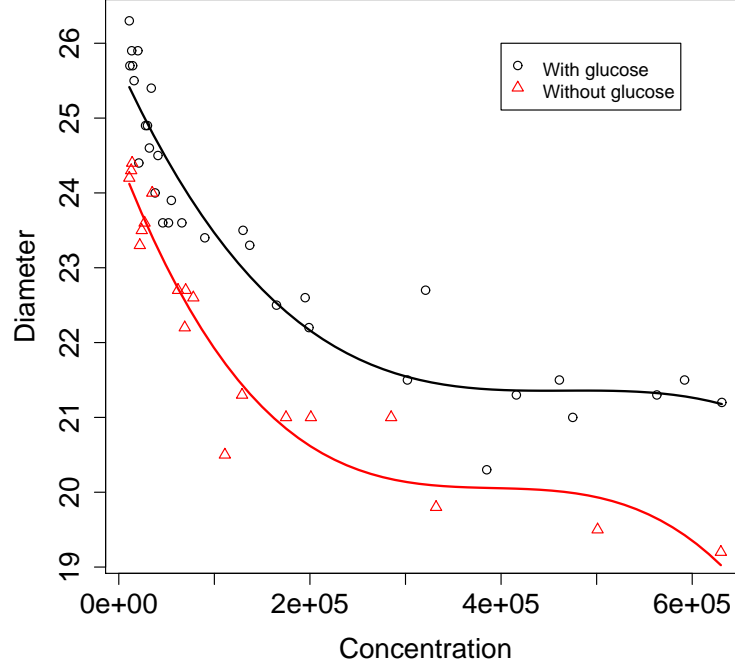


Figure 3: The *hellung* dataset: diameter *vs.* concentration for *Tetrahymena* cells.

402 cell cultures: cells with and without glucose added to the growth medium.  
 403 For each group, the average cell diameter (in  $\mu\text{m}$ ) and the cell concentration  
 404 (count per ml) were recorded. The cell concentrations of both groups were  
 405 set to the same value at the beginning of the experiment and it is expected  
 406 that the presence of glucose in the medium affects the growth of the cell  
 407 diameter. In the sequel, cells with glucose will be considered as coming  
 408 from population  $P$  (32 observations) whereas cells without glucose will be  
 409 considered as coming from population  $P^*$  (between 11 to 19 observations).

410 *Experimental setup.* In order to fit a regression model on the cell group with  
 411 glucose, the PRESS criterion was used to select the most appropriate basis  
 412 function. It results that a 3rd degree polynomial function is the most adapted  
 413 model for these data and this specific basis function will be used for all  
 414 methods in this experiment. Figure 3 shows the ordinary least square (OLS)  
 415 estimates of the 3rd degree polynomial regression model respectively for the  
 416 cell population  $P$  (with glucose) and the cell population  $P^*$  (without glucose).  
 417 The first remark suggested by this figure is that the right extremity of the  
 418 OLS regression curve of population  $P^*$  (bottom red line) is very influenced  
 419 by the last observation. This highlights the non-robustness of this regression  
 420 model learned on only 19 points. The goal of this experiment is to compare  
 421 the stability and the effectiveness of the usual OLS regression method with  
 422 our adaptive linear regression models according to the size of the  $P^*$  learning  
 423 dataset. For this, 4 different learning datasets are used: all  $P^*$  observations  
 424 (19 obs.), all  $P^*$  observations for which the concentration is smaller than  
 425  $4 \times 10^5$  (17 obs.), smaller than  $2 \times 10^5$  (14 obs.) and smaller than  $1 \times 10^5$  (11  
 426 obs.). In order to evaluate the prediction ability of the different methods,  
 427 the PRESS criterion as well as the MSE value on the whole  $P^*$  dataset are  
 428 computed for these 4 different sizes of learning dataset.

429 *Experimental results.* Figure 4 illustrates the effect of the learning set size on  
 430 the prediction ability of the studied regression methods. The panels of Fig-  
 431 ure 4 displays the curve of the usual OLS regression method (M0) in addition  
 432 to the curves of the 5 adaptive linear models (models M1 to M5) for different  
 433 sizes of the learning set (the blue zones indicate the ranges of the observa-  
 434 tions of  $P^*$  used for learning the models). The model M6 which is equivalent

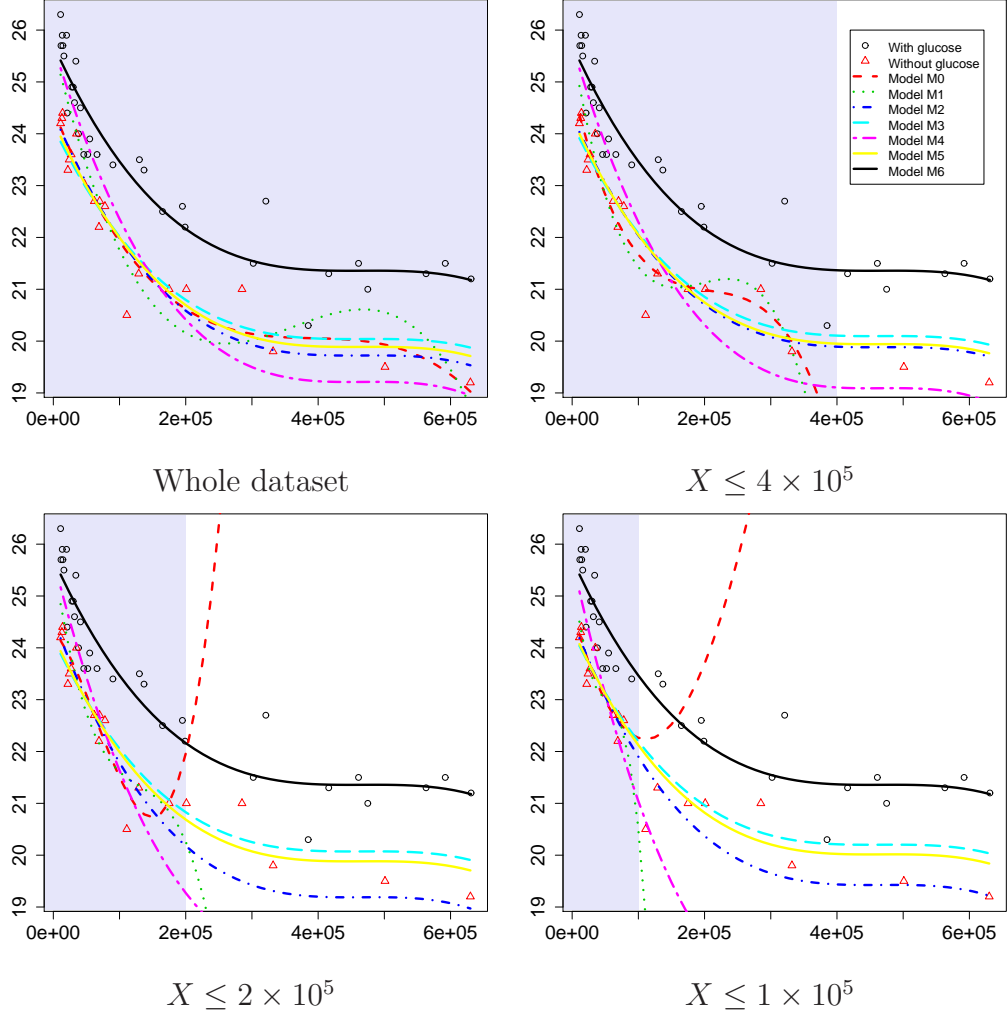


Figure 4: Effect of the learning set size on the prediction ability of the studied regression methods for the *hellung* dataset. The blue zones correspond to the parts of the observations of  $P^*$  used for learning the models.

435 to the usual OLS regression method on the population  $P$  is also displayed.  
 436 The first remark suggested by these results is that the most complex models,  
 437 OLS (M0) and M1, appear to be very unstable in such a situation where the  
 438 number of learning observations is small. Secondly, the model M4 is more  
 439 stable but its main assumption (same intercept as the regression model of  
 440  $P$ ) seems to be an overly strong constraint and stops it from fitting correctly  
 441 the data. Finally, the models M2, M3 and M5 turn out to be very stable  
 442 and flexible enough to correctly model the new population  $P^*$  even with very  
 443 few observations. This visual interpretation of the experiment is confirmed  
 444 by the numerical results presented in Tables 3 and 4. These tables respec-  
 445 tively report the value of the PRESS criterion and the MSE associated to the  
 446 studied regression methods for the different sizes of learning dataset. Table 3  
 447 confirms clearly that the most stable, and therefore appropriate, model for es-  
 448 timating the transformation between populations  $P$  and  $P^*$  is the model M5.  
 449 Another interesting conclusion is that both models M2 and M3 obtained very  
 450 low PRESS values as well. These predictions of the model stability appear  
 451 to be satisfying since the comparison of Tables 3 and 4 shows that the model  
 452 selected by the PRESS criterion is always an efficient model for prediction.  
 453 Indeed, the Table 4 show that the most efficient models in practice are the  
 454 models M2 and M5 which are the “preferred” models by PRESS. These two  
 455 models consider a shift of the intercept, which confirms the guess that we can  
 456 have by examining graphically the dataset, and moreover by quantifying this  
 457 shift. To conclude, this study has shown that the adaptive linear models  
 458 can be successfully applied to real data for transferring a knowledge on a ref-  
 459 erence population (here the cells with glucose) to a new population (here the

Table 3: Effect of the learning set size on the PRESS criterion of the studied regression methods for the *hellung* dataset. The best values of each column are in bold.

| Method            | whole dataset | $X \leq 4 \times 10^5$ | $X \leq 2 \times 10^5$ | $X \leq 1 \times 10^5$ |
|-------------------|---------------|------------------------|------------------------|------------------------|
| OLS on $P^*$ (M0) | 0.897         | 0.364                  | 0.432                  | 0.303                  |
| Model M1          | 3.332         | 0.283                  | 2.245                  | 0.344                  |
| Model M2          | 0.269         | 0.294                  | 0.261                  | 0.130                  |
| Model M3          | 0.287         | 0.271                  | 0.289                  | 0.133                  |
| Model M4          | 0.859         | 1.003                  | 0.756                  | 0.517                  |
| Model M5          | <b>0.256</b>  | <b>0.259</b>           | <b>0.255</b>           | <b>0.124</b>           |

Table 4: Effect of the learning set size on the MSE value of the studied regression methods for the *hellung* dataset. Best values of each column are in bold and the stars indicate the selected models by the PRESS criterion.

| Method            | whole dataset | $X \leq 4 \times 10^5$ | $X \leq 2 \times 10^5$ | $X \leq 1 \times 10^5$ |
|-------------------|---------------|------------------------|------------------------|------------------------|
| OLS on $P^*$ (M0) | 0.195         | 47.718                 | $4.5 \times 10^3$      | 145.846                |
| Model M1          | 0.524         | 164.301                | $2.3 \times 10^3$      | $5.9 \times 10^5$      |
| Model M2          | <b>0.218</b>  | <b>0.226</b>           | 0.304                  | <b>0.245</b>           |
| Model M3          | 0.258         | 0.262                  | 0.259                  | 0.290                  |
| Model M4          | 0.791         | 0.796                  | 1.472                  | 3.046                  |
| Model M5          | *0.230        | *0.233                 | * <b>0.230</b>         | *0.246                 |
| OLS on $P$ (M6)   | 2.388         | 2.388                  | 2.388                  | 2.388                  |

460 cells without glucose). As it could be expected, the advantage of adaptive  
461 linear models makes particularly sense when the number of observations of  
462 the new population is limited and this happens frequently in real situations  
463 due to censorship or to technical constraints (experimental cost, scarcity,...).

### 464 *5.3. Real data study: Modelling of housing market in different U.S. cities*

465 In this section, the interest of the adaptive linear models is illustrated  
466 by an application to the modeling of housing market in different U.S. cities.  
467 This application aims to demonstrate that it is possible to adapt a regression  
468 model learned on a reference city to another one *via* the adaptive linear  
469 models by using only few samples from the new city and, thus, to save an  
470 expensive collect of new data.

471 *The data.* For this experiment, the 1984 American Housing Survey of the  
472 U.S. Department of Commerce is used. The data collection [25] contains  
473 information from samples of housing units in 11 Metropolitan Statistical  
474 Areas, among which the cities of Birmingham, Alabama (East coast) and  
475 of San Jose, California (West coast). Fourteen relevant features have been  
476 selected among more than 500 available features for modeling the housing  
477 market of Birmingham. The selected features include the number of rooms,  
478 the area, the monthly cost of the housing as well as other informations about  
479 the unit and the tenants. Finally, based on these 14 features, the response  
480 variable to predict is the value of the housing.

481 *Experimental setup.* A semi-log regression model for the housing market of  
482 Birmingham was learned using all the 1541 available samples and, then,  
483 the 7 adaptive linear models were used to transfer the regression model of

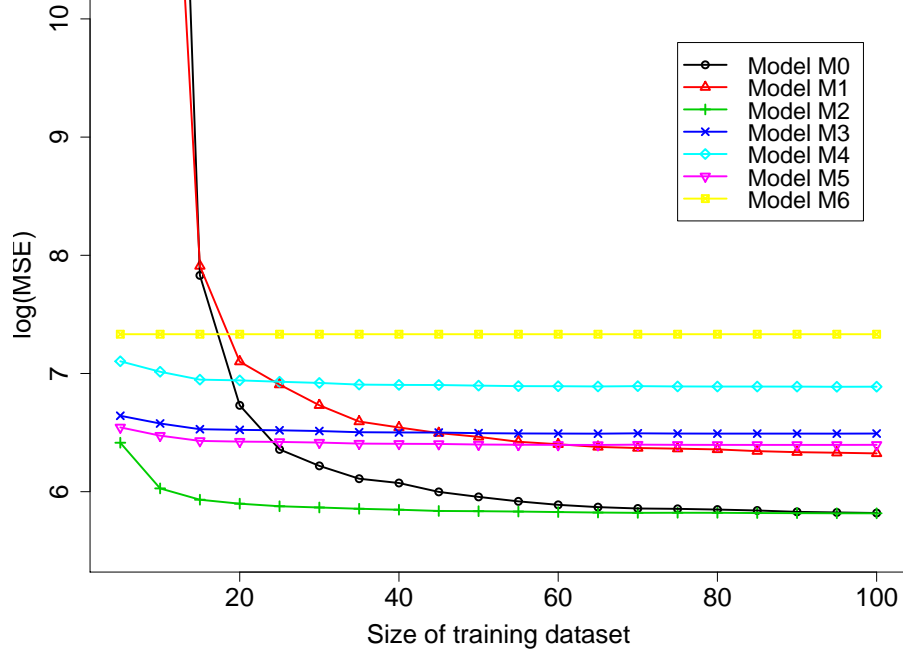


Figure 5: MSE results for the Birmingham-San Jose data.

484 Birmingham to the housing market of San Jose. In order to evaluate the  
 485 ability of the adaptive linear models to transfer the Birmingham knowledge  
 486 to San Jose in different situations, the experiment protocol was applied for  
 487 different sizes of San Jose samples ranging from 5 to 921 observations. For  
 488 each dataset size, the San Jose samples were randomly selected among all  
 489 available samples and the experiment was repeated 50 times for averaging  
 490 the results. For each adaptive linear model, the PRESS criterion and the  
 491 MSE were computed, by using the selected sample for PRESS and the whole  
 492 San Jose dataset for MSE.



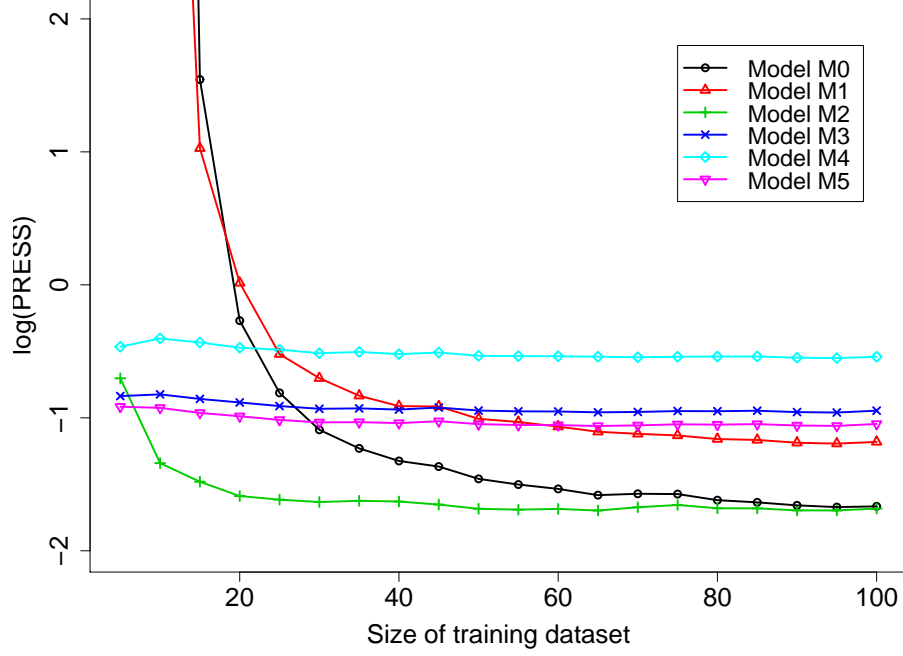


Figure 6: PRESS criterion for the Birmingham-San Jose data.

493 *Experimental results.* Figure 5 shows the logarithm of the MSE for the differ-  
494 ent adaptive linear models regarding to the size of the used San Jose samples.  
495 Similarly, Figure 6 shows the logarithm of the PRESS criterion. Firstly, Fig-  
496 ure 5 indicates that the model M6, which corresponds to the Birmingham's  
497 model, is actually not adapted for modeling the housing market of San Jose  
498 since it obtains a not satisfying MSE value. Let us notice that the curve  
499 corresponding to the MSE of the model M6 is constant since the regression  
500 model has been learned on the Birmingham's data and consequently does  
501 not depend on the size of the San Jose's dataset selected for learning. Sec-  
502 ondly, the model M0, which is equivalent to OLS on the San Jose samples,

503 is particularly disappointing (large values of MSE) if it is learned with a  
 504 very small number of observations and becomes more efficient for learning  
 505 datasets larger than 50 observations. The model M1 has a similar behaviour  
 506 for small learning datasets but turns out to be less interesting than M0 when  
 507 the size of the learning dataset is larger. These behaviours are not surprising  
 508 since both models M0 and M1 are very complex models and then need large  
 509 datasets to be correctly learned. Conversely, the models M2 to M5 appear  
 510 not to be sensitive to the size of the dataset used for adapting the Bir-  
 511 mingham model. Particularly, the model M2 obtains very low MSE values for  
 512 a learning dataset size as low as 20 observations. This indicates that the  
 513 model M2 is able to adapt the Birmingham model to San Jose with only 20  
 514 observations. Moreover Table 5 indicates that the model M2 provides better  
 515 prediction results than the model M0 for the housing market of San Jose for  
 516 learning dataset sizes less than 100 observations. Naturally, since the model  
 517 M0 is more complex, it becomes more efficient than the model M2 for larger  
 518 datasets even though the difference is not so big for large learning datasets.  
 519 Figure 6 shows that the PRESS criterion, which will be used in practice  
 520 since it is computed without a validation dataset, allows the practitioner to  
 521 successfully select the most appropriated transfer model. Indeed, it appears  
 522 clearly that the PRESS curves are very similar to the MSE curves computed  
 523 on the whole dataset. Finally, in such a context, the transformation param-  
 524 eters obtained by the different adaptive linear models can be interpreted in  
 525 an economic way and this could be interesting for economists. In particular,  
 526 the estimated transformation parameters by the model M2 with the whole  
 527 San Jose dataset are  $\lambda_0 = 1.439$  and  $\lambda = 0.447$ . The fact that the San Jose's

Table 5: MSE results for the Birmingham-San Jose data.

| Model    | 10 obs.           | 25 obs.      | 50 obs.      | 100 obs.     | 250 obs.     | all obs.     |
|----------|-------------------|--------------|--------------|--------------|--------------|--------------|
| Model M0 | $3.5 \times 10^7$ | 576.9        | 386.1        | 336.8        | <b>310.7</b> | <b>297.5</b> |
| Model M2 | <b>414.8</b>      | <b>356.7</b> | <b>342.1</b> | <b>336.0</b> | 332.5        | 330.1        |
| Model M6 | 1528.9            | 1528.9       | 1528.9       | 1528.9       | 1528.9       | 1528.9       |

528 intercept is almost 50% larger than the one of Birmingham suggests that  
 529 the minimal basis price of an housing is more expensive in San Jose than in  
 530 Birmingham. However, the fact that the regression coefficients associated to  
 531 the explanatory variables of San Jose are on average 50% smaller than the  
 532 one of Birmingham could mean that the growing of the price according to  
 533 the housing features is more moderated. To summarize, this experiment has  
 534 shown that the adaptive linear models are able to transfer the knowledge  
 535 on the housing market of a reference city to the market of a different city  
 536 with a small number of observations. Furthermore, the interpretation of the  
 537 estimated transformation parameters could help the practitioner to analyse in  
 538 an economic way the differences between the studied populations.

## 539 6. Discussion

540 Before each statistical analysis, the indispensable collect of data is often  
 541 an expensive step. Even though the same analysis has been achieved in a  
 542 relatively similar situation, a new collect of data is needed since the situ-  
 543 ation is usually not exactly similar. In a regression framework, this paper  
 544 shows how it is possible to adapt a regression model from a given situa-  
 545 tion to another new one, and thus to save an expensive new collect of data.

546 In this perspective, a family of adaptive linear models has been introduced  
547 and, since they are more parsimonious than a complete regression model,  
548 they need only few samples for providing satisfying estimation of the new  
549 regression model. To summarize, the main interest of this work arises when  
550 the sample size for the new population is too small to efficiently estimate  
551 a regression model by the usual OLS procedure without using information  
552 known for the reference population. The conducted experiments have shown  
553 that the proposed adaptive linear models are able to successfully transfer a  
554 knowledge on a reference population to another population even with very  
555 few observations. In particular, the efficiency of the proposed models has  
556 been illustrated on a economic application by adapting the regression of the  
557 housing price *versus* housing features from the city of Birmingham to the city  
558 of San Jose. While a sample size of at least 100 observations is needed to  
559 estimate directly the San Jose's regression model, only 20 data are necessary  
560 to obtain a similar estimation quality with the adaptive linear models. In  
561 addition, the estimated transformation parameters could help practitioners to  
562 analyse the differences between both populations. This could be the subject  
563 of a further study and of a collaboration with the economists who provided  
564 these data. Another interesting perspective of this work concerns the pres-  
565 ence of correlation between the covariates. Indeed, if the correlation between  
566 variables is different from one population to the other, it will be necessary  
567 to consider different transformation parameters for these variables.

## 568 Acknowledgments

569 The authors would like to thank Professor Patrice Gaubert (University  
570 Paris XII) for providing the preprocessed economical data and for his very  
571 useful advices and Professor Christophe Biernacki (University Lille I) for  
572 comments and discussions.

## 573 References

- 574 [1] S. R. Searle, Linear models, John Wiley & Sons Inc., New York, 1971.
- 575 [2] P. McCullagh, J. A. Nelder, Generalized linear models, Monographs on  
576 Statistics and Applied Probability, Chapman & Hall, London, 1983.
- 577 [3] D. Ratkowsky, Handbooks of nonlinear regression models, Chapman &  
578 Hall, London, 1990.
- 579 [4] N. Draper, H. Smith, Applied regression analysis, 3rd Edition, John  
580 Wiley & Sons Inc., New York, 1998.
- 581 [5] H. Shimodaira, Improving predictive inference under covariate shift by  
582 weighting the log-likelihood function, J. Statist. Plann. Inference 90 (2)  
583 (2000) 227–244.
- 584 [6] A. Storkey, M. Sugiyama, Mixture regression for covariate shift, Ad-  
585 vances in Neural Information Processing Systems 19, MIT Press, Cam-  
586 bridge, 2007, pp. 1337–1344.
- 587 [7] M. Sugiyama, K.-R. Müller, Input-dependent estimation of generaliza-  
588 tion error under covariate shift, Statistics & Decisions 23 (2005) 249–279.

- 589 [8] M. Sugiyama, K. M. Müller, K-R., Covariate shift adaptation by impor-  
590 tance weighted cross validation, *Journal of Machine Learning Research*  
591 8 (2007) 985–1005.
- 592 [9] M. Sugiyama, Active learning in approximately linear regression based  
593 on conditional expectation of generalization error, *Journal of Machine*  
594 *Learning Research* 7 (2006) 141–166.
- 595 [10] C. Biernacki, F. Beninel, V. Bretagnolle, A generalized discriminant rule  
596 when training population and test population differ on their descriptive  
597 parameters, *Biometrics* 58 (2) (2002) 387–397.
- 598 [11] J. Jacques, C. Biernacki, Extension of model-based classification for  
599 binary data when training and test populations differ, *Journal of Applied*  
600 *Statistics* (2010) in press.
- 601 [12] F. Beninel, C. Biernacki, Modèles d’extension de la régression logis-  
602 tique, *Revue des Nouvelles Technologies de l’Information, Data Mining*  
603 *et apprentissage statistique : application en assurance, banque et mar-*  
604 *keting (A1)* (2007) 207–218.
- 605 [13] N. Feudale, N. Woody, H. Tan, D. Kell, J. Maddock, Heginbothom,  
606 J. M., Magee, Transfer of multivariate calibration models: a review,  
607 *Chemometrics and Intelligent Laboratory System* 64 (2002) 181–192.
- 608 [14] Y. Wang, V. D., K. B., Multivariate instrument standardization, *Ana-*  
609 *lytical chemistry* 63 (23) (1991) 2750–2756.
- 610 [15] R. Goodacre, E. Timmins, A. Jones, D. Kell, J. Maddock, H. M.,

- 611 J. Magee, On mass spectrometer instrument standardization and inter-  
 612 laboratory calibration transfer using neural networks, *Analytica Chim-*  
 613 *ica Acta* 348 (1) (1997) 511–532.
- 614 [16] K. Bertness, R. Hickernell, S. Hays, D. Christensen, Noise reduction in  
 615 optical in situ measurements for molecular beam epitaxy by substrate  
 616 wobble normalization, *Journal of Vacuum Science and Technology B*  
 617 16 (3) (1998) 1492–1497.
- 618 [17] K. Tobin, T. Karnowski, A. L., R. Ferrell, J. Goddard, F. Lakhani,  
 619 Content-based image retrieval for semiconductor process characteriza-  
 620 tion, *Journal on Applied Signal Processing* 1 (2002) 704–713.
- 621 [18] C. Bishop, Pattern recognition and machine learning, *Information Sci-*  
 622 *ence and Statistics*, Springer, New York, 2006.
- 623 [19] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learn-  
 624 ing, *Springer Series in Statistics*, Springer-Verlag, New York, 2001.
- 625 [20] S. Mallat, A wavelet tour of signal processing, 2nd Edition, Academic  
 626 Press, 1999.
- 627 [21] G. Chow, Tests of equality between sets of coefficients in two linear  
 628 regressions, *Econometrica* 28 (1960) 591–605.
- 629 [22] D. M. Allen, The relationship between variable selection and data aug-  
 630 mentation and a method for prediction, *Technometrics* 16 (1974) 125–  
 631 127.

- 632 [23] G. Schwarz, Estimating the dimension of a model, The Annals of Statis-  
633 tics 6 (2) (1978) 461–464.
- 634 [24] H. Akaike, A new look at the statistical model identification, IEEE  
635 Transactions on Automatic Control 19 (6) (1974) 716–723.
- 636 [25] W. D. United States Department of Commerce, Bureau of the Census,  
637 American housing survey, 1984: Msa file, Ann Arbor, MI: Inter- univer-  
638 sity Consortium for Political and Social Research.